

HARALD ANDERSON

EFFICIENCY VERSUS PROTECTION
IN
RANDOMIZED RESPONSE DESIGNS



LUND 1975



EFFICIENCY VERSUS PROTECTION

IN

RANDOMIZED RESPONSE DESIGNS

av

Harald Anderson

civ.ing., Klm

Akademisk avhandling som för avläggande av teknisk
doktorsexamen vid tekniska fakulteten vid universitetet
i Lund kommer att offentligen försvaras å matematiska
institutionen, Sölvegatan 18, sal A, torsdagen den 11
september 1975 kl. 10.15.



Harald Anderson

EFFICIENCY VERSUS PROTECTION
IN
RANDOMIZED RESPONSE DESIGNS

Lund 1975



Printed in Sweden

AV-centralen i Lund

Lund 1975

ACKNOWLEDGEMENTS

Quite a few people have been of great help to me in accomplishing this thesis.

I wish to express my gratitude to professor Gunnar Blom for his encouraging interest in my work and for his valuable advice concerning both content and form of the thesis;

Dr Jan Lanke, who by numerous discussions throughout the work has supported me very much and who carefully read the manuscripts pointing out both linguistic and other errors.

I am also indebted to my other colleagues and friends at the departments of Mathematical Statistics and Mathematics, University of Lund, for the open and generous atmosphere at our institute. In particular I would like to mention Dr Lennart Bondesson and FK Daniel Thorburn with whom I have had many constructive discussions and who have read and commented on earlier versions of the papers.

Finally I want to thank professor Tore Dalenius, who some years ago directed my attention to the field of survey sampling and who always has shown interest in my work.

Lund, July 1975

Harald Anderson

The thesis consists of this summary and the following two papers:

[A] Anderson, H.: Efficiency versus protection in randomized response designs for estimating proportions. Report 1975:9, Department of Mathematical Statistics, University of Lund.

[B] Anderson, H.: Efficiency versus protection in a general randomized response model. Report 1975:10, Department of Mathematical Statistics, University of Lund.

1. INTRODUCTION

During the last decade people have become more concerned than before about delivering information about themselves for survey purposes as well as for government and other registers. This change in attitude is a consequence of the increasing amount of data collected and of the rapid development of data processing: people fear that the information might be used against them in some way or another. To cope with the problems caused by this concern, statisticians, computer specialists and legislators are taking an increased interest in "data security", i.e. in how to protect data from being misused.

Two fundamentally different situations can be distinguished when considering the purpose of the collection of data:

- (a) one is primarily interested in data from single individuals,
- (b) one is only interested in aggregated data, e.g. population means and standard deviations.

Situation (a) occurs for instance when the assessment authority wants to know everybody's income and situation (b) occurs for instance when a sampling survey is performed for predicting the result of an election. In situations of type (a) it is quite clear that all individuals must provide information on themselves. The collected data can be protected only by rigorous precautions against improper use. In situations of type (b) the information can be protected already at the assembling stage. To begin with, it is not necessary, although often preferable, to label the data with the names of the individuals; the population values can still be estimated. However, very often, the interviewers will know the characteristics of the individuals interviewed, or, the interviewees may be identified by means of the data provided, and these facts may lead to evasive answers or answer refusal.

This thesis is devoted to an interviewing technique which admits

unbiased estimates of the population parameters without knowledge of the true value of any single individual. The technique is called randomized response (RR for short), and it was introduced by Warner (1965) as a means of reducing answer bias when dealing with embarrassing subjects.

Randomized response is based on a very simple idea. Suppose one wants to estimate the proportion π of a population that has some sensitive characteristic A. The sampled persons are asked to draw and answer a question from a pack of cards with two types of questions,

Are you an A-individual?,

and

Are you an A*-individual?,

A* denoting not A. The proportions of cards are p and $1-p$, respectively, where p is a number $\neq 1/2$ determined by the statistician. Since the interviewer does not know which question is answered, he cannot infer the attribute of the respondent. The proportion π can still be estimated, because the probability of a yes-answer is a linear function of π .

Since 1965 much work has been done on RR; new RR-schemes for estimating proportions and means of a population have been proposed, new randomizing devices have been constructed and experience has been gained on how people react when confronted with RR.

The core of RR is a randomizing device, by means of which the respondent gives an answer, partly random and partly dependent on the value of his characteristic. The idea is that, from these answers, one should be able to obtain estimates of the population parameters without direct knowledge of the sampled individuals' characteristics.

The aim of an RR-plan is twofold:

- (i) to afford good estimates of the population parameters
- and

(ii) to attain a sufficient degree of protection to the interviewees.

The randomizing device should be chosen so as to satisfy these requirements. This leads to a variety of practical and theoretical problems. Hitherto most research work on RR has dealt with questions concerning the construction and test of different RR-schemes in order to decrease the answer bias and answer refusal in special situations. Then the design of the randomizing mechanism and the behaviour of the interviewer play important roles. This approach to RR is very natural and essential from a practical point of view.

In this thesis are considered some theoretical aspects of RR, which make it possible to insert the RR-techniques in a wider statistical context. The first paper (paper [A]) deals with the problem of estimating proportions by means of RR, starting with the single-proportion case. The second paper (paper [B]) is more general and covers the estimation of any population parameters.

We shall first discuss the single-proportion case, which gives the basic ideas of the thesis.

2. ESTIMATION OF A SINGLE PROPORTION BY MEANS OF RR

Suppose that the population is divided into two classes, A and A*, in unknown proportions π and $1-\pi$, respectively. The following model is proposed:

A-individuals give the random answer Y according to the known probability density $f_0(y)$,

and

A*-individuals give the random answer Y according to the probability density $f_1(y)$.

This model covers every RR-scheme where π is the only unknown parameter. For example it includes the well-known schemes of Warner and Simmons and also schemes with repeated interviews (Y may be multi-dimensional). Moreover, every pair of response densities

$f_0(y)$ and $f_1(y)$ can be realized, for instance by a spinner with two scales, one for A- and one for A*-individuals.

The observable answer Y from a randomly taken individual has the density

$$(1) \quad f_Y(y; \pi) = \pi f_0(y) + (1-\pi)f_1(y), \quad 0 < \pi < 1.$$

Hence π is a mixture parameter between two known densities, $f_0(y)$ and $f_1(y)$.

To design an RR-plan for estimating a proportion π is, from a statistical point of view, equivalent to choosing the two response densities $f_0(y)$ and $f_1(y)$. According to the aims (i) and (ii) above, which are basically conflicting, two aspects must then be considered: the plan should provide an acceptable protection to the interviewees and the plan should allow an estimate of π with a sufficiently small variance.

Let us first consider the efficiency of a plan to estimate π . In surveys the sample size is generally large. Then the Cramér-Rao bound (CR-bound for short) yields a satisfactory approximation of the variance of a "good estimate" π^* of π . Moreover, using this bound we can make comparisons between different designs, $(f_0(y), f_1(y))$, without specifying any estimator. Fisher's information about π is

$$I = E\left[\frac{\partial}{\partial \pi} \log f_Y(Y; \pi)\right]^2 = E\left[\frac{f_0(Y) - f_1(Y)}{f_Y(Y; \pi)}\right]^2,$$

and the CR-bound based on n independent observations of Y is $1/nI$. It is immediately seen that I is small and hence the CR-bound large if the response densities $f_0(y)$ and $f_1(y)$ are much alike.

A measure of how well an RR-plan protects the privacy of the respondents is more difficult to construct. In most papers the protection is discussed in terms of the parameters of the special design used, e.g. the probability p of answering the question "Are you an A-individual?" in Warner's scheme. We shall use another approach based on an idea of Lanke (1975a,b). The conditional probabilities of being an A-individual and an A*-individual given the

answer $Y=y$ are

$$(2) \quad \begin{aligned} P(A|y;\pi) &= \frac{\pi f_0(y)}{\pi f_0(y) + (1-\pi) f_1(y)} , \\ P(A^*|y;\pi) &= \frac{(1-\pi) f_1(y)}{\pi f_0(y) + (1-\pi) f_1(y)} . \end{aligned}$$

These probabilities, called risks of suspicion, are used for measuring the protection associated with the answer $Y=y$. The closer $P(A|y;\pi)$ is to π , and hence $P(A^*|y;\pi)$ to $1-\pi$, the larger is the protection.

Lanke (1975b) and independently Leysieffer (1975) have used the above conditional probabilities when comparing the efficiency of RR-schemes with two possible answers, yes and no, for estimating a proportion π (e.g. Warner's and Simmons's schemes).

In paper [A] it is shown that the variance of the maximum-likelihood estimate $\hat{\pi}$ based on n independent observations can be written

$$V(\hat{\pi}) = \frac{\pi(1-\pi)}{n} \cdot \frac{1}{\left[\frac{P(A|\text{yes};\pi)}{\pi} - 1 \right] \cdot \left[\frac{P(A^*|\text{no};\pi)}{1-\pi} - 1 \right]} .$$

Hence, for a two-answer plan, there is a simple relation between the variance of the best estimator and the protection as expressed by the risks of suspicion.

For a general plan $(f_0(y), f_1(y))$ Fisher's information can be rewritten, using (2), in the form

$$I = E(P(A|Y;\pi) - \pi)^2 / \pi^2 (1-\pi)^2 = V(P(A|Y;\pi)) / \pi^2 (1-\pi)^2 .$$

Both in the two-answer case and in the general case we see that the better the respondents are protected the less efficient is the plan to estimate π .

The above discussion summarizes the qualities of RR-schemes with different response distributions. How should a statistician plan an RR-survey, i.e. how should these distributions be chosen? In paper [A] it is suggested that $f_0(y)$ and $f_1(y)$ be taken so that, for a fixed π ,

$$(3) \quad k_1 \leq P(A|y;\pi) \leq k_2$$

for every possible answer y and some constants k_1 and k_2 . If the characteristic A is regarded as embarrassing, k_2 should be chosen smaller than one and if A^* is regarded as embarrassing, k_1 should be chosen larger than zero. The unknown proportion π must be guessed on the basis of apriori information. Condition (3) guarantees the respondent a minimum protection whatever his answer will be.

If the guess about π is wrong, the risks of suspicion might be changed. Therefore we also discuss the possibility of requiring that the condition (3) should hold for every π in some interval (π_1, π_2) believed to contain the true value.

Subject to Condition (3) (both for a fixed π and all $\pi \in (\pi_1, \pi_2)$) it is shown that a plan with two possible answers minimizes the variance of the "best" π -estimate. Hence an optimal plan is determined by the two response probabilities $P(\text{yes}|A\text{-individual})$ and $P(\text{yes}|A^*\text{-individual})$. It is easily seen that any pair of response probabilities can be generated using Simmons's unrelated question model.

3. ESTIMATION OF SEVERAL PROPORTIONS BY MEANS OF RR

Let the population consist of the classes $A_1, \dots, A_k$, in the proportions $\pi_1, \dots, \pi_k$, $\sum \pi_i = 1$, and let $f_i(y)$ be the known response density from A_i -individuals, $i = 1, \dots, k$. The answer from a randomly taken individual then has the density

$$f_Y(y; \pi) = \pi_1 f_1(y) + \dots + \pi_k f_k(y).$$

As in the single-proportion case the efficiency of an RR-plan to estimate $\pi = (\pi_1, \dots, \pi_k)^T$ is measured by the corresponding information matrix. The protection associated with the answer $Y=y$ is measured by the vector $(P(A_1|y; \pi), \dots, P(A_k|y; \pi))$, where

$$P(A_i|y; \pi) = \pi_i f_i(y) / \sum \pi_i f_i(y)$$

is the conditional probability that the respondent is an A_i -individual if he has answered $Y=y$.

The analogue to Condition (3) now becomes more complex. We have to prescribe a $(k-1)$ -dimensional set in which $P(A|Y;\pi) = (P(A_1|Y;\pi), \dots, P(A_{k-1}|Y;\pi))^T$ is to take values for some guessed value of π . This set, Q_π , is called a risk restriction set. The response densities $f_i(y)$, $i = 1, \dots, k$, should be taken so that the random vector $P(A|Y;\pi)$ with probability one takes values within Q_π . Such plans are called Q_π -plans.

Since π is multi-dimensional, there is no single plan that is optimal for estimating the whole vector π . Considering the set of all linear combinations $\lambda = \sum t_i \pi_i$ of π , a class of admissible Q_π -plans is introduced. A Q_π -plan $(f_1(y), \dots, f_k(y))$ with information matrix (I_{ij}) is called admissible if there does not exist another Q_π -plan with information matrix (I'_{ij}) such that

$$\sum t_i t_j I_{ij} \leq \sum t_i t_j I'_{ij}$$

for all $t_1, \dots, t_{k-1}$, with strong equality for some $t_1, \dots, t_{k-1}$.

If Q_π is a convex simplex, it is shown that the random vector $P(A|Y;\pi)$ for an admissible Q_π -plan with probability one takes values in the vertices of Q_π . Moreover, the optimal Q_π -plan for estimating a specified linear combination $\lambda = \sum t_i \pi_i$ has at most $2k-2$ possible answers.

At the end of paper [A] proportions depending on a parameter θ are treated. This situation occurs for instance when a distribution $G(x;\theta)$ is partitioned into classes A_i with $\pi_i = \int_{A_i} dG(x;\theta)$. When θ is one-dimensional, determination of the optimal Q_π -plan leads to a simple linear-programming problem.

4. A GENERAL RR-MODEL

Suppose that the variable X is distributed according to the density $f_X(x;\theta)$, where $\theta = (\theta_1, \dots, \theta_r)^T \in \Theta$ is unknown. Using some random mechanism an individual with true value $X=x$ gives the answer Y according to the known probability density $h_Y(y|x)$, $x \in \Omega_x$; here Ω_x is the set of possible x -values. The family of

response densities $\{h_Y(y|x); x \in \Omega_X\}$ determines the statistical properties of the RR-scheme used. The density of the answer from a randomly taken individual is

$$f_Y(y; \theta) = \int h_Y(y|x) f_X(x; \theta) dx = E(h_Y(y|X)),$$

and the protection associated with the answer $Y=y$ is given by the conditional density of X given $Y=y$,

$$h_X(x|y; \theta) = h_Y(y|x) f_X(x; \theta) / f_Y(y; \theta).$$

Let us assume that θ is one-dimensional and that the family $\{f_X(x; \theta); \theta \in \Theta\}$ fulfils necessary regularity conditions. Then Fisher's information from a direct observation of X is

$$I(X) = E \left[\frac{\partial \log f_X(X; \theta)}{\partial \theta} \right]^2.$$

On the other hand, if the randomized response Y is observed, Fisher's information is

$$I(Y) = E \left[\frac{\partial \log f_Y(Y; \theta)}{\partial \theta} \right]^2.$$

In paper B it is shown that

$$I(X) - I(Y) = I(X|Y),$$

where $I(X|Y)$ is the expected conditional information from X given Y ,

$$I(X|Y) = E_Y E_X \left[\frac{\partial \log h_X(X|Y; \theta)}{\partial \theta} \right]^2.$$

Hence the loss of information $I(X|Y)$ due to RR is directly related to the protection of the respondent's privacy as expressed by the family of revealing densities $\{h_X(x|y; \theta); y \in \Omega_Y\}$, Ω_Y being the set of possible RR-answers.

In the general case it is difficult to give explicit recommendations on the design of an RR-plan, since the demands of protection differ very much from situation to situation. However, a minimal protection for every possible answer y is advocated (as in Condition (3)), and in two examples some possible conditions are suggested.

Paper [B] also treats the problem of testing two simple hypotheses by means of RR. The loss of power due to RR for a fixed sig-

nificance level is estimated by means of Kullback-Leibler's information. Analogously to the estimation case the loss of information equals the expected conditional Kullback-Leibler information from X given Y .

The last section of paper [B] deals with two examples. In the first of these $\{f_X(x;\theta); \theta \in \Theta\}$ is assumed to be an exponential family of densities, and in the second a special RR-model, "the model with given answers" (Eriksson (1973)) is treated. In both examples the aim is to illustrate the conflict between efficiency and protection.

REFERENCES

- Eriksson, S. (1973). Randomized Interviews for Sensitive Questions. Ph.D. Thesis, Institute of Statistics, University of Göteborg.
- Lanke, J. (1975a). On the choice of the unrelated question in Simmons' version of randomized response. J. Amer. Statist. Assoc. 70, 80-83.
- Lanke, J. (1975b). On the degree of protection in randomized interviews. Report no. 2 of the research project Confidentiality in Surveys, Institute of Statistics, University of Stockholm.
- Leysieffer, F.W. (1975). Respondent jeopardy in randomized response procedures. Technical report, Department of Statistics, Florida State University.
- Warner, S.L. (1965). Randomized response: A technique for eliminating evasive answer bias. J. Amer. Statist. Assoc. 60, 63-69.

